Methodological Issues in Analyzing Real-World Longitudinal Occupational Health Data: A Useful Guide to Approaching the Topic

Rémi Colin-Chevalier, Frédéric Dutheil, Sébastien Cambier, Samuel Dewavrin, Thomas Cornet, Julien Steven Baker, Bruno Pereira

International Journal of
Environmental Research and Public Health

MDPI

Review

Methodological Issues in Analyzing Real-World Longitudinal Occupational Health Data: A Useful Guide to Approaching the Topic

Rémi Colin-Chevalier [1,2,3,*], Frédéric Dutheil [1,2,3], Sébastien Cambier [4], Samuel Dewavrin [3], Thomas Cornet [3], Julien Steven Baker [5] and Bruno Pereira [4]

1 CNRS, LaPSCo, Physiological and Psychosocial Stress, Université Clermont Auvergne, F-63000 Clermont-Ferrand, France; frederic.dutheil@uca.fr
2 Preventive and Occupational Medicine, CHU Clermont-Ferrand, F-63000 Clermont-Ferrand, France
3 Wittyfit, F-75000 Paris, France; samuel.dewavrin@wittyfit.com (S.D.); thomas.cornet@wittyfit.com (T.C.)
4 Biostatistics Unit, The Clinical Research and Innovation Direction, University Hospital of Clermont-Ferrand, CHU Clermont-Ferrand, F-63000 Clermont-Ferrand, France; scambier@chu-clermontferrand.fr (S.C.); bpereira@chu-clermontferrand.fr (B.P.)
5 Centre for Health and Exercise Science Research, Hong Kong Baptist University, Kowloon Tong, Hong Kong 999077, China; jsbaker@hkbu.edu.hk
* Correspondence: r.colin6374@gmail.com

check for updates

Citation: Colin-Chevalier, R.; Dutheil, F.; Cambier, S.; Dewavrin, S.; Cornet, T.; Baker, J.S.; Pereira, B. Methodological Issues in Analyzing Real-World Longitudinal Occupational Health Data: A Useful Guide to Approaching the Topic. *Int. J. Environ. Res. Public Health* **2022**, *19*, 7023. https://doi.org/10.3390/ijerph19127023

Academic Editors: Vasile Palade, Alireza Daneshkhah, Amin Hosseinian-Far and Samer A. Kharroubi

Received: 17 May 2022
Accepted: 6 June 2022
Published: 8 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract: Ever greater technological advances and democratization of digital tools such as computers and smartphones offer researchers new possibilities to collect large amounts of health data in order to conduct clinical research. Such data, called real-world data, appears to be a perfect complement to traditional randomized clinical trials and has become more important in health decisions. Due to its longitudinal nature, real-world data is subject to specific and well-known methodological issues, namely issues with the analysis of cluster-correlated data, missing data and longitudinal data itself. These concepts have been widely discussed in the literature and many methods and solutions have been proposed to cope with these issues. As examples, mixed and trajectory models have been developed to explore longitudinal data sets, imputation methods can resolve missing data issues, and multilevel models facilitate the treatment of cluster-correlated data. Nevertheless, the analysis of real-world longitudinal occupational health data remains difficult, especially when the methodological challenges overlap. The purpose of this article is to present various solutions developed in the literature to deal with cluster-correlated data, missing data and longitudinal data, sometimes overlapped, in an occupational health context. The novelty and usefulness of our approach is supported by a step-by-step search strategy and an example from the Wittyfit database, which is an epidemiological database of occupational health data. Therefore, we hope that this article will facilitate the work of researchers in the field and improve the accuracy of future studies.

Keywords: methodological issues; modeling; occupational health; longitudinal data; missing data; cluster-correlated data; real-world data

1. Introduction

In previous decades, randomized clinical trials have been the main experimental methodology used to collect clinical data. Collected within a strict framework, the analysis of randomized clinical trials data is used to answer specific questions related to a specific population (namely the population selected for the study). This remains true provided that analyses are performed rigorously with consideration for all required hypotheses and methods (e.g., definition of a representative population sample, randomization, controlled tests, application of appropriate models to the data). However, the advances in technology, the democratization of computer tools such as computers and smartphones as well as the increase in data storage capacities offer researchers new possibilities for collecting

large amounts of health-related data for analysis. Such data, collected independently of traditional trials, is also known as real-world data [1–3]. More broadly, the data can be obtained from various sources such as patient health records, product and disease registers or even digital health data collection platforms and applications. Many examples of real-world databases can be listed and classified according to the sources from which they come. As examples of this, we can cite records from patient registries such as the European Cystic Fibrosis Society Registry database [4], healthcare databases such as Wittyfit and Wittyfit Research [5], pharmacy and health insurance databases such as the Food and Drug Administration's Sentinel Initiative [6], social media such as PatientsLikeMe [7] or patient-powered research networks such as PCORnet [8]. Big data management approaches such as the management of chronic kidney disease [9], chronic obstructive pulmonary disease [10], or lymphoma subtypes [11] have naturally emerged from such databases.

As a perfect complement to randomized clinical trials data and presenting a longitudinal structure [12], the results of the analysis of real-world data, also called real-world evidence, are very popular; for example, within pharmacology research, these data sets are considered essential because they are crucial in pharmacological decision-making [13,14]. However, by construction, real-world databases are not immune to missing, noisy (outliers), duplicate or inconsistent data phenomena [15].

A longitudinal study is a study investigating repeated data, i.e., where the same data has been measured/collected on an individual several times over time [16]. Unlike cross-sectional studies, cohort effects and time-related effects can be measured separately. The main goal of a longitudinal study is to describe changes over time and measure the individual influence of variables to explain changes observed [17]. Longitudinal data may also facilitate the change over time when investigating particular individuals [18]. For example, it is possible to measure risk factors in the development of disease for particular individuals in the population [19]. However, many challenges and methodological issues make it difficult to analyze longitudinal data. These analytical problems include the correlated structure of intra-individual data, the considerable size of data sets, irregular time-spaced measurements, non-linear patterns (such as rapid growth or stationary responses), latent constructs, and the mix of time-varying and static covariates.

There are also more difficult problems to consider, such as cluster-correlated data, missing data and longitudinal modeling itself [20,21]. To sum up, studying such data types requires understanding the specific concepts that are outlined in Figure 1.

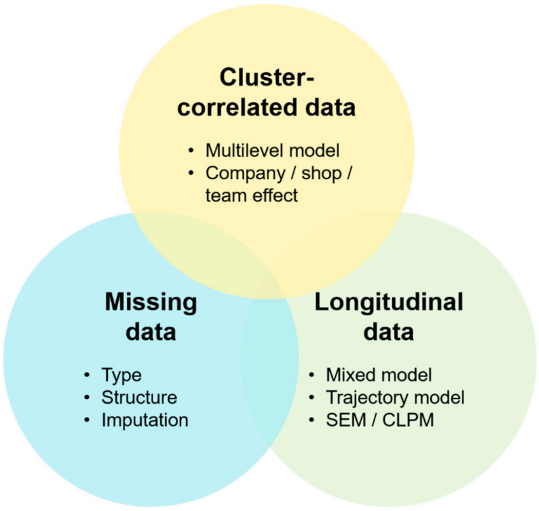


Figure 1. Data diversity of real-world longitudinal occupational health databases (see Appendix A for the novelty and usefulness of our approach and main formulas surrounding cluster-correlated data, missing data, and longitudinal data). SEM: structural equation model, CLPM: "cross-lagged" panel model.

Although a multitude of approaches have been proposed in the literature, it is still difficult to account for all these concerns in a synchronized way. The purpose of this paper is to review the state of the art of different methods and models designed to address issues related to real-world longitudinal occupational health data in a holistic way to improve the quality of future surveys on the topic. In the first part, we aim to introduce each methodological issue and to present advanced methods for addressing them. In the second part, we aim to present a case study in which all the issues overlap and present a method to meet the objectives of the study using the solutions previously featured.

2. Methodological Issues and Methods for the Analysis of Longitudinal Data

2.1. Cluster-Correlated Data

Cluster-correlated data occur when pre-existing groups of individuals exist within the population, allowing for a natural classification of subjects into groups, otherwise known as clusters. Thus, the observations of the same cluster are correlated with each other while the observations between the different clusters are uncorrelated. This structure requires considering two possible sources of variability, namely intra- and inter-cluster variability. These similarities in the data can, however, lead to a loss of statistical efficiency and integrity of the models used, necessitating an increase in the size of the sample [22]. It is therefore necessary to find an appropriate balance between the number of clusters and the number of patients.

Correlated data occur in a number of different settings. They can be clustered, spatial (grouping by location), multilevel or longitudinal data or even several at the same time [23]. A hierarchical or multilevel structure can be thought of as a series of several levels nested within each other. Thus, the lowest level represents level 1, and the levels follow each other upwards to the level where all the data is contained. The units formed at each level can then be likened to clusters. A hierarchical structure is therefore a succession of more and more specific clusters and, from observation of the lower levels, is a structure that is "neither accidental nor ignorable" [24]. Multilevel structures widen the field of possibilities in terms of questions that can be answered (e.g., presence of a group effect, disparities within the different clusters). On the other hand, the structures question the use of statistical methods that do not consider the nature of the structures, which could also provide erroneous results [25,26]. Longitudinal data is a special form of clustered data. By nature, an individual's data is plausibly more positively correlated with each other than with other individuals. Like correlated data, this intra-subject correlation must be taken into account in the analyses, otherwise this can provide false positive results and erroneous confidence intervals [21,27]. Figure 2 presents an example of a population divided into clusters followed in a longitudinal study.

Thus, cluster-correlated data, regardless of its type, requires special analytical treatment [28]. Cluster-correlated data analysis attempts to take into account the variability associated with each level of the structure and must be interpreted separately from the overall variability [29]. Indeed, misspecification of the cluster effect or careless interpretation of the model parameters can lead to erroneous results [30,31]. Hence, in the literature, multilevel models (MLMs) have been designed to address these various issues and analyze correlated-clustered data [24,29,32,33]. These models are similar to mixed models. In fact, the "levels" described by the multilevel models can be assimilated to be the "random effects" of the mixed model. This implies that cluster-correlated data can be analyzed properly using the models defined above by considering the levels of the data as random effects. Similarly, for longitudinal data, this implies that the individual effect must be considered as a random effect also.

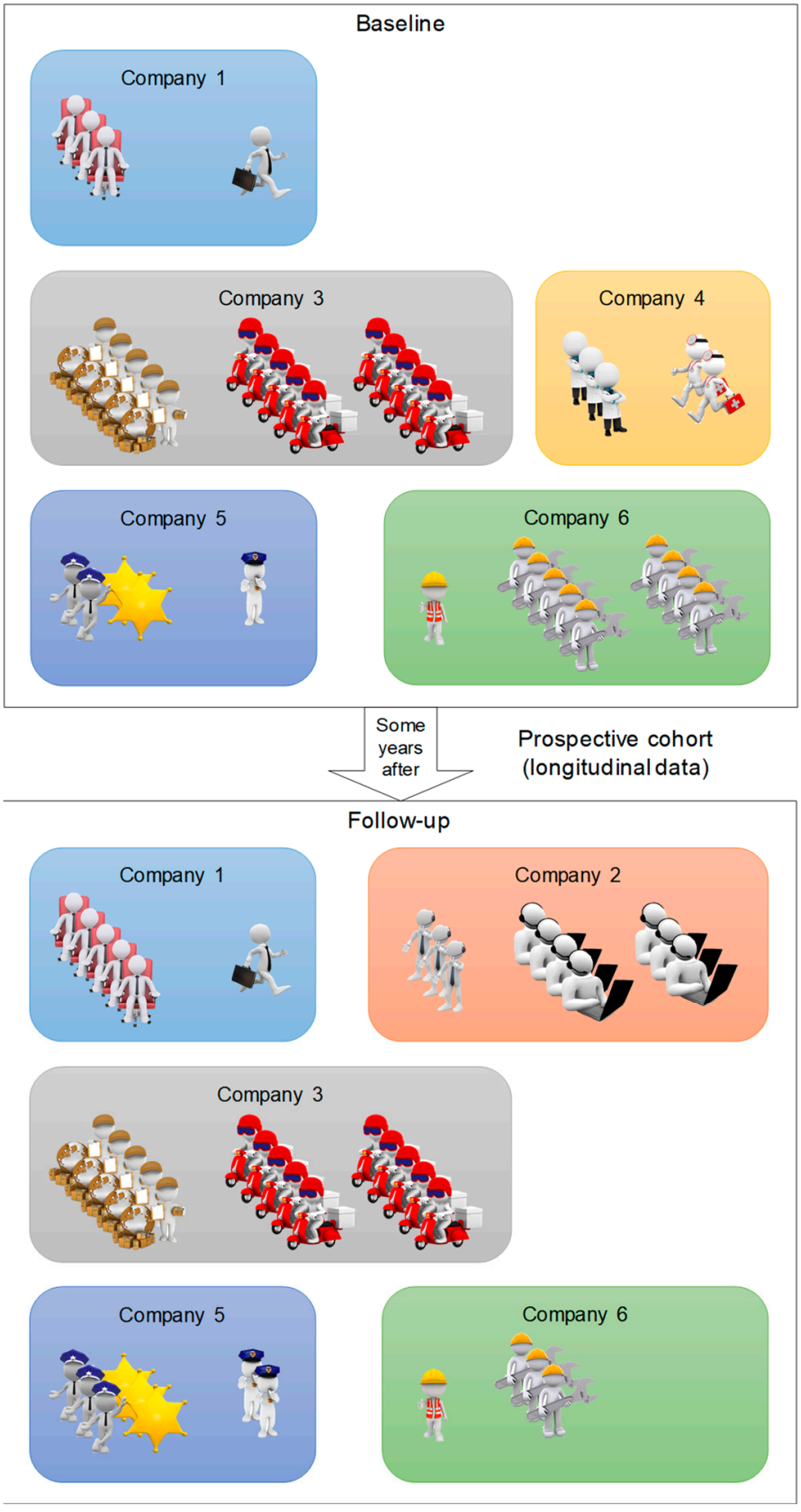


Figure 2. Example of cluster-correlated data: here, a prospective cohort of workers from multiple companies is followed over time.

Clustered data are not immune to the phenomenon of missing data. In the absence of data, conventional imputation methods can be applied. However, the particular structure of cluster-correlated data, which leads to multiple sources of variability (intra- and inter-cluster), suggests proceeding differently. As a result, it appears preferable to apply the imputation method chosen cluster by cluster [34]. To sum up, in the presence of cluster-correlated data, it is preferable to estimate the missing values within clusters. When the data of an individual is missing, estimates can be made using individuals from the same cluster. For a hierarchical structure, the lowest level cluster in which the individual is located will be preferred. If an entire cluster leaves the study, the missing data can be estimated from the data of individuals at the level above. Finally, at the highest level of the hierarchy, it can be estimated using the entire data set.

2.2. Missing Data

The phenomenon of missing data is a common problem in data analysis, especially in longitudinal surveys. It is rare for a data set to be complete, particularly for long studies, either because of occasional omissions or because of subject withdrawal. As a consequence, data missingness causes three main issues [35]: it can introduce a significant amount of bias, make it more difficult to process and analyze the data and induce a loss of statistical power [36,37]. When faced with missing data, the most common used methods include the complete case analysis (or listwise deletion) and the "last observation carried forward" methods. However, these methods are too hazardous and can introduce a significant bias in the estimation of the parameter under investigation [38–40]. As previously observed, these ad hoc methods are no longer desirable or necessary. Likewise, in the presence of clustered data, it is difficult to estimate what effect the loss of an individual or even an entire cluster has on the outcome measures. Accordingly, it is better not to ignore missing values but to try to estimate them using imputation methods. The purpose of these methods is to estimate missing values from previously observed values, which can be considered as measured data and therefore used for fitting analytical models. While these methods may seem attractive, their use is not without problems [41]. Even though there are some methods that are more effective than others, it is not really possible to say which method is especially good [42]. These methods require understanding and apprehension of certain concepts that we aim to outline below.

To be more effective and valid, statistical analysis must be performed with suitable mechanisms and assumptions for missing data [37]. In other words, it is crucial to understand the process behind it to ensure the validity of the statistical inferences and the absence of bias [43]. Therefore, a careful preparation should be observed before using a method that is very similar to an identification process. In particular, it is essential to identify the two main parameters: the structure and the type of missing data.

The structure plays an important role in the choice of imputation methods that can be applied [44]. There are three possible structures, which can overlap. The structure is said to be univariate if one and only one variable of an individual has missing values. It is said to be monotonic when several variables of an individual have missing values and these variables can be classified according to their percentage. For longitudinal surveys, this structure can be found when an individual leaves the study, for example in case of attrition or even abandonment. Attrition, or random drop-out [45], is a major and common problem in longitudinal studies, representing an important challenge when modeling [46] as it may produce estimates of the effects that are often underestimated [47]. If it is not possible to obtain the data of an individual at a given time (e.g., forgotten appointment, forgotten connection), we often consider intermittent monotonic structures. The structures can qualify as arbitrary (or non-monotonic) if they are neither univariate nor monotonic. Figure 3 illustrates the different structures of missing data.

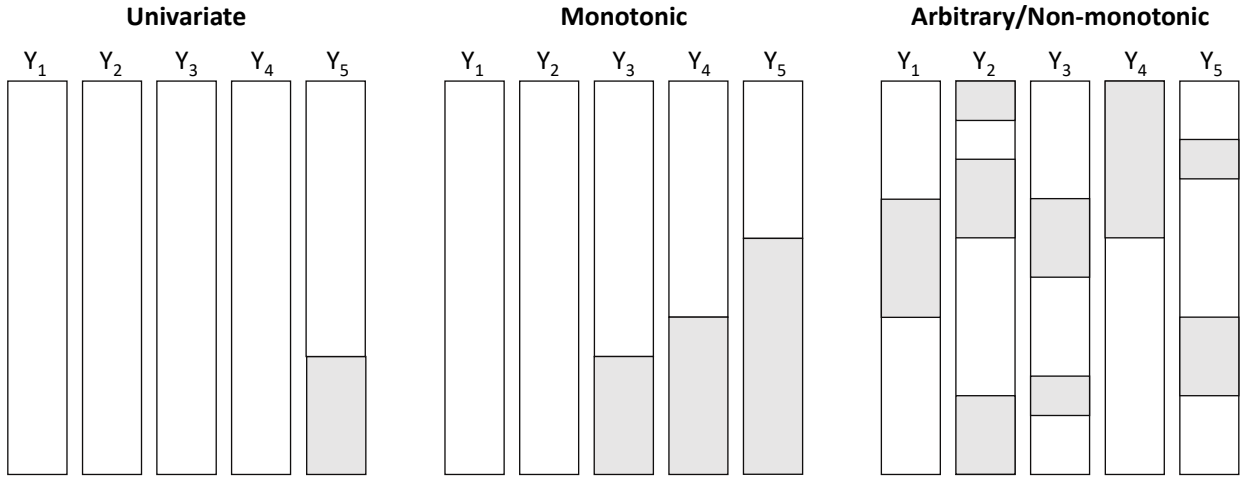


Figure 3. Missing data structures. White sections indicate the presence of data, shaded sections their absence.

After identifying the structure of missing data, it is necessary to deal with the identification of its type. We consider here again three different types: data which is missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [41]. A missing value is considered to be MCAR if its missingness does not depend on other observations, whether they are observed or not. Under this assumption, the results of the analyses appear to be generally unbiased, but it is rarely verified because it is restrictive. A less restrictive assumption supposes that if it depends on observed observations only, its type will be MAR. Otherwise, if it is neither MCAR nor MAR, it will be MNAR. Despite this formulation, it is not always possible in practice to identify the causes of data missingness and therefore to determine its type [48].

Finally, once the type and structure of the missing data are identified, an imputation method can be used. Table 1 provides a non-exhaustive overview of the different methods that can be used to calculate estimates of missing values [41,49].

Table 1. Imputation methods by type of missing data.

Missing Completely at Random		Missing at Random	Missing Not at Random
Ad hoc methods			
Complete case analysis, available-case analysis, weighting methods		Expectation maximization algorithm	"Sensitivity analysis"
Single imputation			
Implicit modeling	**Explicit modeling**		
Hot/cold deck imputation, substitution, composite methods	Mean/regression/stochastic regression imputation	Multiple imputation	

For the particular case of MNAR type data, it is necessary to perform a "sensitivity analysis". Indeed, although there is a multitude of models to analyze this type of data, they all rely on assumptions (e.g., the missing data mechanism) which, for this type of missing data, are unverifiable. When the results are particularly sensitive to the assumptions made, it becomes difficult to choose the most suitable model. That is, instead of fitting a single model, it is wiser to consider alternative models and to assess the sensitivity of the results to the assumptions made about the missing data mechanism [31,41,48].

2.3. Longitudinal Data and Modeling

Repeated data models have been the subject of many years of research and development. Today, the literature on the subject is substantial and many of these models have been used in longitudinal surveys. Among the models used most for the analysis for longitudinal data [20,50], the most often used are variance models, whether univariate (ANOVA) or multivariate (MANOVA), random effects models such as mixed models or

generalized linear models (GLMs) using generalized estimating equations. Nevertheless, other models such as trajectory models or alternatives to linear and mixed models such as structural equation models (SEMs) or "cross-lagged" panel models (CLPMs) have appeared in the literature and constitute promising alternatives for the analysis of longitudinal data. These models are not all similar and are intended to answer different questions. While mixed models and their alternatives are used to estimate the impact of factors on a response variable, trajectory models are used to model individual trajectories within a population and follow their evolution over time. Given a large amount of data, modeling should be performed on a training sample and model validation on a test sample. We will now describe the above-mentioned models in more detail.

2.3.1. Analysis of Variance for Repeated Measures

The analysis of variance for repeated measures represents a classical method to analyze longitudinal data. Whether for ANOVA or MANOVA, they allow comparing groups' means on a dependent variable across time. However, it does not allow learning about individual trajectories. In addition to other parameters, time is also treated as an explicative variable. This means, among other things, that these models assume that individuals cannot have a proper slope over time, which is rarely true. Moreover, these models are difficult to apply in the absence of data and are applied using the complete cases or the "last observation carried forward" methods. As a consequence, it is therefore best to avoid these models and to turn to more suitable methods for repeated data analysis [27]. For example, mixed models offer more advantages [36], especially by adding a subject-specific component to the model and allowing the conduct of the analysis despite possible lack of data.

2.3.2. Mixed Models

Mixed models can be seen as natural extensions of regression models [51,52]. Unlike the latter, they involve random effects specific to each individual. The mean and the variance of the response are respectively modeled as a linear combination of fixed-effect and random-effect components, where the impact of the factors is weighted and associated with the coefficients of the model [53,54]. Unlike analysis of variance models, both marginal and mixed models are unconditional. This feature makes it possible to model the response variable as a function of both the covariates and time separately representing both within- and between-subject effects [20]. They are also particularly effective for making individual predictions even despite eventual missing data because the randomly missing estimates are unbiased [41]. More broadly, generalized linear mixed models combine the specificities of generalized linear models (GLMs) and mixed models, allowing for generalizing the type of the explained variable. This change in construction is ideal for the analysis of longitudinal data structures since it makes it possible to account for the intra-subject correlation through random effects [55,56].

This makes mixed models a safe and effective method for the analysis of longitudinal data. Other more specific models, as we will see below, have been designed subsequently and added to the literature, but mixed models remain the most popular. Once a trajectory is identified, it is possible to describe it over time but also in terms of static covariates using a mixed model or equivalent.

2.3.3. Generalized Estimating Equations

In general, the estimation of GLM parameters is based on the maximum likelihood method [57]. Yet, the generalized estimating equations method does represent a popular alternative to likelihood-based generalized linear mixed models, allowing the model to be extended to the analysis of correlated data. Afterwards, they simultaneously model the link between explicative variables, the response and the within-subject dependence. This method is particularly appreciated on the one hand because it gives efficient and unbiased estimates of the parameters of a generalized linear model and on the other hand because it accounts for intra-individual correlation [58,59]. Generalized estimating equations can deal

with missing data under the assumptions that they are MCAR. When these data are MAR, the estimates might be biased [60].

2.3.4. SEMs and CLPMs, Complementary Approaches to the Mixed Model

Despite its effectiveness, the mixed model is only a first step towards more complex statistical methods allowing the move from a global analysis of the population to a more personalized/individualized analysis of occupational data. Therefore, models such as SEMs or CLPMs can be used as alternatives to mixed models for the analysis of longitudinal data.

SEMs can be viewed as a much more comprehensive regression method, including dependent and independent variables. Where SEMs stand out is in the ability to additionally take into account hypothetical latent constructs, and to examine relationships between observed variables and these concepts [61–65]. As such, they can be seen as a natural combination between factor analysis and regression or path analysis. Although they are particularly powerful, SEMs stay very sensitive to the problem of missing values. In addition to the assumptions of normality and independence that they require, SEMs need a very large sample size to fit. On average, at least 500 individuals and up to >2500 individuals if one of the assumptions is not verified are needed to expect good estimates [65].

CLPMs are used to describe and estimate reciprocal relationships or directional influences between longitudinal variables [66,67]. Cross-analysis is particularly used to describe causal relationships between variables. Often contested because of their low statistical power and their limitations (e.g., the need for a large sample of longitudinal data, the stationarity and synchronicity assumptions or the causal relationship assertion), simpler methods such as multiple regression are often preferred [68]. Because of this, alternative models to SEMs such as the random intercept "cross-lagged" panel models have been developed to overcome their shortcomings [69]. In addition to the temporal stability assumption of the classical method, these consider the trait-type stability invariant of individuals over time.

2.3.5. Trajectory Models

By plotting longitudinal data, we obtain curves otherwise called individual trajectories or developmental trajectories. A developmental trajectory describes and provides information on the evolution of an individual over time. A trajectory is used to describe a latent process that cannot be observed directly but can be explained over time using measured (therefore observed) variables from which its trajectory is inferred [70,71]. These methods that take into account unobserved heterogeneity are called person-centered [72]. Nowadays, several common methods [73,74] are used: generalizations of mixed effects models such as growth curve models [75], growth mixture models [76,77], group-based trajectory models [78–81], latent class analysis [82,83] and latent transition analysis [72].

Most of these methods identify several different trajectories within the population, also called latent classes. For each trajectory, the estimated share of individuals in the population belonging to it is given, which reflects its shape, and each individual has a certain probability of belonging to each trajectory. Individuals are then assigned to the group corresponding to the trajectory for which the probability of belonging is the highest. Once a trajectory is identified, it is possible to explain its shape over time but also in terms of static covariates using a mixed model or equivalent.

Although an individual trajectory can be estimated despite prospective missing values, it is difficult to conclude on the validity of this trajectory, especially when data is NMAR [84]. It is essential to have a sufficient amount of data for an individual, otherwise it is impossible to determine its trajectory [85]. These models are also considerably sensitive to assumptions, and misspecification can lead to biased estimates for trajectories and an overestimation of the number of classes by the model [73].

3. Case Report

3.1. Introduction

We aim here to illustrate our suggestions with a complete and illustrated example using the Wittyfit database [5]. Here, we focus on a small part of the approaches described above, namely the contribution of trajectory models to mixed models in a real-world longitudinal occupational health data framework.

Let us imagine that we want to analyze the annual evolution in job satisfaction of workers at different companies between 2018 and 2021, and to observe if the gender and the job position of a worker can affect job satisfaction. A first look at the objective allows us to confirm that we are in the presence of clustered (multiple companies), correlated (individual effect) and longitudinal (follow-up over time) data. If a worker did not express his or her sentiment in a year, he or she is assigned a missing value. Thus, we are faced with the main methodological issues mentioned in this paper.

3.2. Methods

3.2.1. Participants and Exlusion Criteria

Wittyfit software is a web-based platform designed to assess workers' health through a holistic approach of the individual. Volunteers are invited to express their feelings on different health-related outcomes using visual analog scales. With more than 40,000 active users in about nearly 80 companies, whose first registrations began in January 2018, the Wittyfit database offers researchers a substantial behavioral and longitudinal database to study the evolution of workers' occupational health through various indicators such as job satisfaction and stress. Workers not present at the baseline (2018) or with too few data [85] were excluded.

3.2.2. Outcomes

Job satisfaction of workers was assessed using a related visual analog scale, scaled from 0 to 100. Workers could rate their personal feeling of job satisfaction as many times as they wanted. A worker's overall annual job satisfaction score was computed as the average of the notes that he or she filled in over the year.

Socio-demographic characteristics of workers (gender and job position) were filled in by corporates clients of Wittyfit. The job position of a worker is defined according to whether he is an employee or a manager.

3.2.3. Statistics

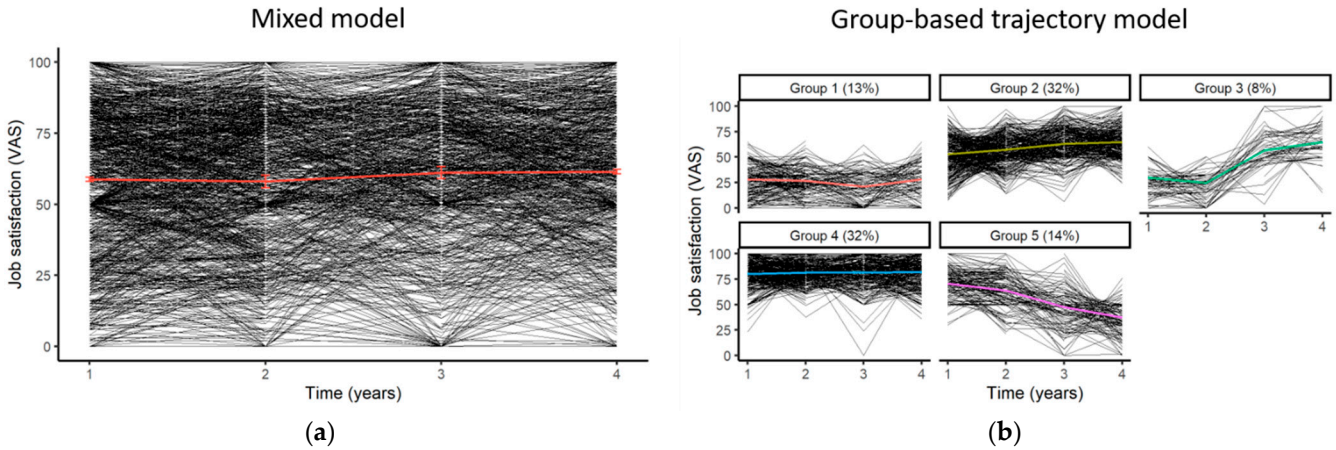
Statistical analyses were performed using R (version 4.1.1) in the RStudio (version 1.4.1717) platform. The imputation of the data was done using "longitudinalData" (with the 'locf' and 'linear.interpol' methods) and "mice" packages. Group-based trajectory modeling was realized with the "latrend" package. We combine the commands 'lcMethodLcmmGBTM' and 'lcMethods' to define the model and the 'latrendBatch' command to fit it, with nonstructured matrix of variance-covariance. Model assumptions (residual independence and normality, and variance homogeneity) were verified a posteriori. Unless specified, we considered a p-value < 0.05 as statistically significant for analyses.

3.3. Data Application

First, we need to analyze the data structure to determine if it has a multilevel structure in addition to a longitudinal structure itself. In our example, we count multiple companies to which workers belong. Therefore, this observation indicates to consider the companies as clusters. We now need to focus on missing data by analyzing its type and structure. The missing data presents two different patterns: (1) a non-monotonic (e.g., due to a worker who did not express in a year) and (2) a non-monotonic pattern (e.g., due to a worker who did stop expressing over the years). For the second one, as data is longitudinal, a drop-out phenomenon may occur. Since it is reasonable to assume that an individual's drop-out may depend on unobserved data (e.g., for an individual who quit during the studied period

because of a lack of job satisfaction), we need to suppose that the drop-out is informative, and therefore that the type of missing values is MNAR. This assumption should therefore lead us to perform a sensitivity analysis by applying several imputation methods, for example with the "last observation carried forward" and the linear regression interpolation methods, but also using multiple imputation, taking into account the company effect.

To study the evolution of workers' job satisfaction, we can first apply a linear mixed model assessing the influence of time on job satisfaction (Figure 4a). Starting with a single trajectory given by the mixed model, we can observe whether different evolutions exist in the population. To do this, we can apply a trajectory model such as the GBTM on the original dataset and confirm the results using the imputed datasets. The model allows us to identify five latent classes within the population, with a minimal average posterior probability of assignment of 70.5% and minimal odds of correct classification of 5.5, thus exceeding the classical thresholds of 70% and 5 expected for this type of model [79] (Figure 4b).



(**a**) (**b**)

Figure 4. Evolutionary trajectories of job satisfaction among Wittyfit users: (**a**) mean trajectory using a mixed model (standard errors of coefficient are symbolized by error bars), (**b**) individual trajectories using a group-based trajectory model, where each group represent a possible evolution of the workers' job satisfaction in the population (e.g., individuals belonging to "Group 3" are characterized by a slight decrease between times 1 and 2, then by a sharp increase beyond time 2).

Finally, to explore the relationship between a worker's sociodemographic characteristics and job satisfaction, we can apply a generalized mixed model with the trajectory to which the worker belongs as the outcome and the company effect as a random effect. Thus, according to the model, both gender ($\beta = 1.60$, 95% CI 1.15 to 2.24, $p = 0.005$) and job position affect the job satisfaction of the worker ($\beta = 5.49$, 95% CI 2.72 to 13.13, $p < 0.001$). This result remains true regardless of the dataset, non-imputed or imputed.

3.4. Conclusions

From a self-imposed context, we presented here the many challenges we had to face in analyzing real-world longitudinal occupational health data. This study had two objectives, namely the identification of the existence of different evolutionary trajectories of job satisfaction within a population of workers and the role of the job position on the level of job satisfaction. Each methodological issue has been raised and addressed using the appropriate methods in agreement with the purpose of our article, allowing us to meet the aims of the study.

4. Conclusions

As a direct result of the rise of massive databases [5–8] and advances in computational and digital tools and their democratization, real-world data analysis has allowed researchers to confirm the results of previous randomized clinical trials and provide new knowledge, known as real-world evidence [1,3]. The progressive and incessant accumulation

of data thus offers the possibility of collecting health-related, longitudinal, individual data, all at a low cost, thus allowing the development of data-analysis-centered medical approaches.

Although the analysis of longitudinal data has become widespread over the last 30 years, particularly through mixed models, and has made it possible to study population evolution, this article demonstrates that methodological issues remain and have begun to find an echo in more recent approaches, allowing for a move towards a more predictive, preventive, personalized and participatory medicine. Despite the advances offered by these new approaches, some of these issues remain to be addressed before proceeding with the analyses [20,21], in addition to the usual issues linked to any data analysis, limiting their use and overall understanding [74].

In this article, we aimed to present the state of the art of the methods developed for the analysis of real-world longitudinal occupational health data while exposing the three main issues encountered during the analyses of such type of data, namely the concepts of cluster-correlated data, missing data and longitudinal data itself. In addition, we have provided an example of data analysis presenting these issues and discussed the steps to be taken in the analysis process. A search strategy using numerous general article databases was conducted, showing that there is currently no approach to deal with these three methodological issues.

We believe this article can serve as a practical guide for future studies on the topic to improve experimental quality, especially for non-statistician researchers who wish to keep abreast of the new approaches available and the issues involved. Future studies will focus on the comparison between the different models and approaches presented here. Simulation work has already been done [86–89], but not in this specific framework.

Author Contributions: Conceptualization, R.C.-C., B.P. and F.D.; methodology, R.C.-C. and B.P.; software, R.C.-C.; validation, B.P. and F.D.; formal analysis, R.C.-C.; investigation, F.D., S.D. and T.C.; resources, S.D. and T.C.; data curation, R.C.-C.; writing—original draft preparation, R.C.-C.; writing—review and editing, B.P., F.D., J.S.B. and S.C.; visualization, R.C.-C.; supervision, B.P. and F.D.; project administration, B.P.; funding acquisition, S.D. and T.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was approved by the National Commission for Data Protection and Liberties (CNIL), and the South-East VI ethics committee (clinicaltrials.gov identification number NCT02596737).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. This information is set out in Wittyfit's terms and conditions of use.

Data Availability Statement: Data from Wittyfit cannot be transmitted without the prior consent of the company's corporate clients, except to the University Hospital of Clermont-Ferrand, France, which may use the data for research purposes.

Acknowledgments: We express our sincere gratitude to all voluntary workers using Wittyfit, who participated in this study.

Conflicts of Interest: R.C., S.D. and T.C. are part of Wittyfit. Other authors have declared that no competing interests exist (F.D. is responsible for the scientific accuracy of Wittyfit, but is not paid by Wittyfit; as previously published, Wittyfit is a public private partnership with the CHU Clermont-Ferrand).

Appendix A. Novelty and Usefulness of Our Approach and Main Formulas Surrounding Cluster-Correlated Data, Missing Data and Longitudinal Data

Many challenges and methodological issues make it difficult to analyze longitudinal data. These analytical problems include the correlated structure of intra-individual data, the considerable size of data sets, irregular time-spaced measurements, non-linear patterns (such as rapid growth or stationary responses), latent constructs, mix of time-varying and static covariates. There are also more difficult problems to consider, such as longitudinal

modeling itself, missing data and cluster-correlated data. To sum up, studying such data types requires understanding the specific concepts that are outlined in Figure 1.

The novelty and usefulness of our approach is validated by a step-by-step search strategy. The search presented below has been computed on 8 February 2022 using the PubMed database (details of other research can be found in Appendix B). First, we computed the number of articles dealing with occupational data using keywords ("occupation*" OR "profession*" OR "job-related" OR "work-related") followed by one of the three concepts: correlated-clustered data using keyword ("cluster*"), missing data using keywords ("missing data" OR "missing value*") and longitudinal data using keyword ("longitudinal"). We retrieved n = 11,214 articles using the keywords ("occupation*" OR "profession*" OR "job-related" OR "work-related") AND ("cluster*"), n = 809 articles using the keywords ("occupation*" OR "profession*" OR "job-related" OR "work-related") AND ("missing data" OR "missing value*") and n = 19,436 articles using the keywords ("occupation*" OR "profession*" OR "job-related" OR "work-related") AND ("longitudinal"). Then, we combined the two concepts. We retrieved n = 31 articles using the keywords ("occupation*" OR "profession*" OR "job-related" OR "work-related") AND ("cluster*") AND ("missing data" OR "missing value*"), n = 335 articles using the keywords ("occupation*" OR "profession*" OR "job-related" OR "work-related") AND ("cluster*") AND ("longitudinal") and n = 71 articles using the keywords ("occupation*" OR "profession*" OR "job-related" OR "work-related") AND ("missing data" OR "missing value*") AND ("longitudinal"). Lastly, we found only n = 3 articles using the keywords ("occupation*" OR "profession*" OR "job-related" OR "work-related") AND ("cluster*") AND ("missing data" OR "missing value*") AND ("longitudinal"). Therefore, there is currently no approach to deal with the methodological issues in analyzing real-world longitudinal occupational health data. The results of our research can be summarized as shown in Figure A1.

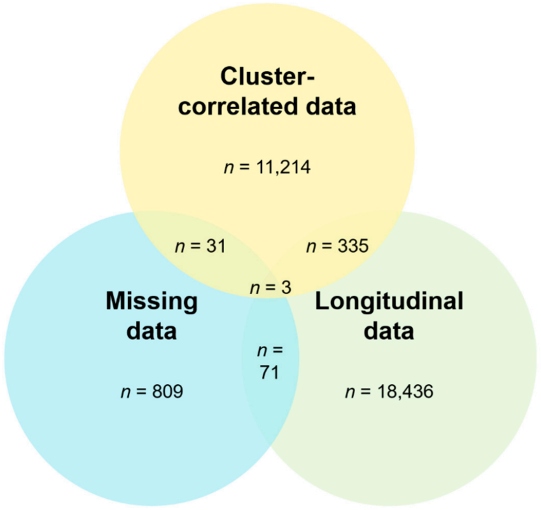


Figure A1. Results of the step-by-step search strategy using PubMed.

To date, the understanding of occupational health data is still under research. As mentioned above, there are three main issues in analyses, namely cluster-correlated data, missing data and longitudinal data itself. Details of the main formulas surrounding cluster-correlated data, missing data and longitudinal data can be found below (Figure A2).

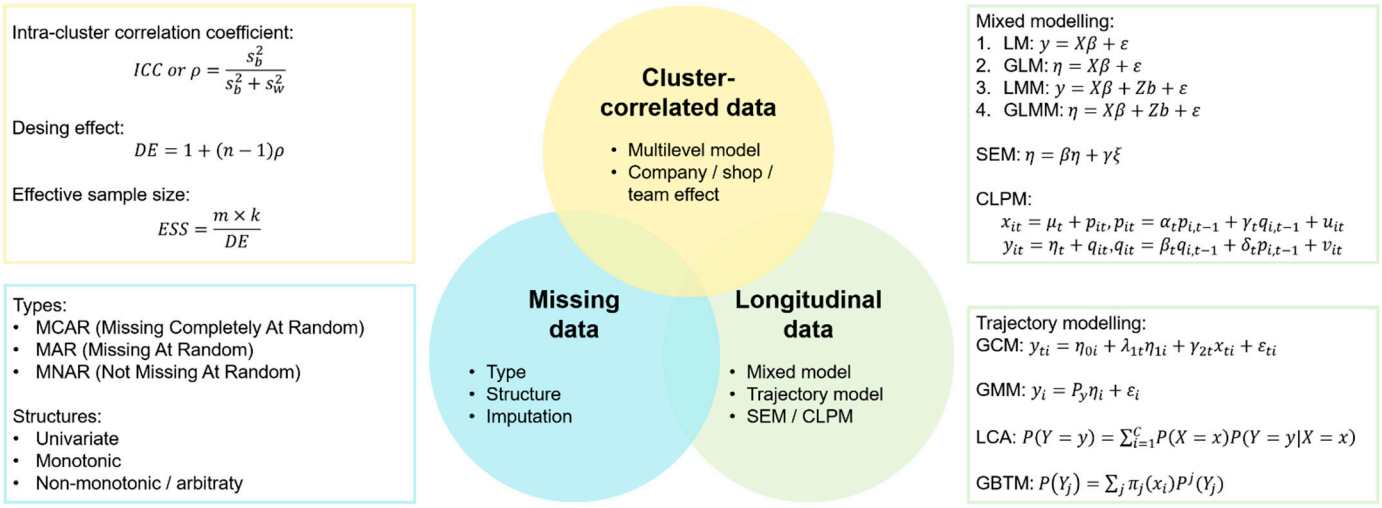


Figure A2. Main formulas for the methods presented to address methodological issues for analyzing real-world longitudinal occupational health data. Table A1 provides a summary of the main models and formulas presented here.

Main Formulas Surrounding Cluster-Correlated Data, Missing Data, and Longitudinal Data.

Appendix A.1. Cluster-Correlated Data

The intra-cluster correlation coefficient (ICC) is a measure of similarity of cluster-correlated data. The ICC can be defined as follows:

$$\text{ICC } or\ \rho = \frac{s_b^2}{s_b^2 + s_w^2} \tag{A1}$$

where s_b^2 represents the between-cluster variability and s_w^2 the within-cluster variability. As the ICC increases, the sample size required to detect a significant effect becomes larger. The "design effect" (DE) estimator is then used to estimate the increase in sample size needed to account for the homogeneity of the clustered data:

$$\text{DE} = 1 + (n-1)\rho \tag{A2}$$

where n is the average size of a cluster. Finally, the "effective sample size" (ESS) can be calculated as bellow:

$$\text{ESS} = \frac{m \times k}{\text{DE}} \tag{A3}$$

where m corresponds to the number of subjects in a cluster and k the total number of clusters.

Appendix A.2. Longitudinal Data

Appendix A.2.1. Linear Modeling

The linear model describes the relationship between one or more independent variables X_i $(i = 1, \cdots, n)$, also called the predictors and a continuous dependent variable y, also called the response. The equation of a linear model is:

$$y = X\beta + \varepsilon \tag{A4}$$

where X represents the matrix of the predictors X_i, β the vector of fixed effects and ε the vector of errors terms following a normal distribution with mean zero and variance $\sigma^2 I_n$.

The generalized linear model is a generalization of the linear model, including a linking function specifying the relationship between the predictors and the response [90]. The equation of a generalized linear model then is:

$$\eta = X\beta + \varepsilon \tag{A5}$$

where η represent the link function. In the logistic regression case, the link function will be the *logit* function described as follows:

$$\eta = logit(p) = log\left(\frac{p}{1-p}\right) \tag{A6}$$

with $p = P(y = 1)$.

The linear mixed model and generalized mixed model appear as extensions of the linear model and generalized mixed model, respectively. Mixed models contain both fixed effects (as describe in the two first models) and random effects [20,53]. The equation of a generalized mixed linear model is:

$$\eta = X\beta + Zb + \varepsilon \tag{A7}$$

where b is the vector of random effects and Z the design matrix relating b to y.

Although likelihood-based approaches are still the most commonly used to estimate model parameters, there are alternatives such as Bayesian approaches [57], generalized estimating equations [58,59] or Monte Carlo-based methods [91].

Appendix A.2.2. Structural Equation Modeling

The structural equation modeling is a set of techniques allowing the estimation of the relationships between dependent and independent variables. Factually, the structural equation models (SEMs) are, as the generalized model, a generalization of the linear model. As such, the equation of a structural equation model seems very close to that of the linear model. It can, be described as follows [61,63]:

$$\eta = \beta\eta + \gamma\xi \tag{A8}$$

where η is the vector of dependent variables, β the matrix of the regression coefficients between dependent variables, γ the matrix of regression coefficients between dependent and independent variables and ξ the vector of independent variables.

Maximum likelihood, generalized least squares, elliptical distribution theory [92] or the asymptotically distribution free method [93] represent possible estimation methods for this model. The choice of the parameter estimation depends on the choice of the weight matrix [65].

Appendix A.2.3. Cross-Lagged Panel Modeling

Cross-lagged panel models (CLPMs) can be useful to measure the evolution of several variables and evaluate their mutual influences over time. Considering two variables x and y measured at multiple times (i.e., $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$), the measurement equations between the two variables for an individual i are:

$$x_{it} = \mu_t + p_{it}y_{it} = \eta_t + q_{it} \tag{A9}$$

where μ_t and η_t are the means of the two variables at time t, respectively and where the temporal deviations are defined as:

$$p_{it} = \alpha_t p_{i,t-1} + \gamma_t q_{i,t-1} + u_{it}q_{it} = \beta_t q_{i,t-1} + \delta_t p_{i,t-1} + v_{it} \tag{A10}$$

where α_t and β_t represent the auto-regressive effects, i.e., the effect of a variable on itself, and γ_t and δ_t the cross-lagged effects, i.e., the effect on a variable on the other, the whole over time.

As for the SEMs, the estimation of the parameters can be computed using the maximum likelihood method [94]. Nevertheless, Bayesian estimation can also represent an alternative to the first method [95].

Appendix A.3. Trajectory Modeling

Appendix A.3.1. Growth Curve Modeling

Trajectory models consist of statistical methods used to analyze the change over time among a population of individuals. The growth curve model (GCM) allows in particular analyzing both inter- (also called the between-person) and intra-individual (the within-person) changes thanks to the mean and the covariance structure, respectively. As well, considering a response vector y measured t times, the model can be written as [75]:

$$y_{ti} = \eta_{0i} + \lambda_{1t}\eta_{1i} + \gamma_{2t}x_{ti} + \varepsilon_{ti} \tag{A11}$$

where λ_{1t} is the measurement of the variable at the first time, the individual intercept η_{0i} with mean v_0 and random departure ξ_{0i} is defined by:

$$\eta_{0i} = v_0 + \gamma_0\zeta_i + \xi_{0i} \tag{A12}$$

and the individual slope with mean v_1 and random departure ξ_{1i} by:

$$\eta_{1i} = v_1 + \gamma_1\zeta_i + \xi_{1i} \tag{A13}$$

Finally, γ_0 and γ_1 represent the effects of the time-invariant covariate ζ_i and γ_{2t} of the time-varying covariate x_{ti} and ε_{ti} the time-specific deviation, following a normal distribution with mean zero and variance σ^2_ε.

Appendix A.3.2. Growth Mixture Modeling

The growth mixture model (GMM) is quite similar to the GCM, with the exception that it combines the techniques of the growth models with the latent concepts defined in the latent class analysis. Considering an observed response y and a set of latent continuous variables η, the model can be written as [76]:

$$y_i = P_y\eta_i + \varepsilon_i \tag{A14}$$

where P_y describes the matrix of the parameters and ε the vector of residuals with mean zero and covariance D. Continuous latent variables η are for them related to the observed covariate x and the latent categorical variables c by the relation:

$$\eta_i = Lc_i + Q_\eta x_i + \xi_i \tag{A15}$$

where L describes the matrix of the intercept parameters of latent classes c, Q_η describes the matrix of the parameters and ξ the vector of residuals with mean zero and covariance D.

Appendix A.3.3. Latent Class Analysis

Latent class analysis assumes the existence of several underlying latent classes in a population. Although the actual class membership of an individual is unknown, it can be inferred through the covariates. The basic equation of the LCA is [82]:

$$P(Y = y) = \sum_{i=1}^{C} P(X = x)P(Y = y|X = x) \tag{A16}$$

where $P(X = x)$ is the probability of an individual belonging to latent class c among the C existing latent classes.

Appendix A.3.4. Group-Based Trajectory Modeling

The estimation of the GBTM parameters is quite similar to the estimation of the LCA parameters. Indeed, the equation of the GBTM is:

$$P(Y_j) = \sum_j \pi_j(x_i) P^j(Y_j) \quad \text{(A17)}$$

where $P^j(Y_j)$ is the probability of Y_j (i.e., the trajectory of the individual i) if the subject belongs to the group j and $\pi_j(x_i)$ is the membership's probability of the group j, determined from the covariate x_i.

Appendix A.4. Summary of Models Formula

Summary of the main formulas surrounding models for cluster-correlated data, missing data and longitudinal data can be found below (Table A1).

Table A1. Summary of the different models.

Model	Search Strategy	Mathematical Formulation	Missing Data	Advantages	Drawbacks
MULTILEVEL MODELING					
MLM	3585	$y_{ij} = \beta_0 + \beta_1 x_{ij} + (u_{0j} + u_{1j} x_{ij} + e_{0ij})$	*a*	- Calculation of the intra-cluster variability apart from the overall variability	- Validity of conclusions highly dependent on the cluster effect specification and definition and interpretation or the model parameters
FIRST APPROACHES					
ANOVA for repeated measures	1547	$SS_{error} = SS_{total} - SS_{between} - SS_{subject} (SS : Sum\ of\ Squares)$	*a*	- Comparison of group's means across time	- No information on possible individual trajectories - Time effect as fixed effect
Mixed model	1983	$y = X\beta + \varepsilon$ $\eta = X\beta + \varepsilon$ $\eta = X\beta + Zb + \varepsilon$	*b* *	- Combination of fixed and random effects - Randomly missing estimates unbiased - Possible consideration of subject-specific/cluster effect	- Only provides information on the average trajectory followed by the population
TRAJECTORIES					
GCM	126	$y_{ti} = \eta_{0i} + \lambda_{1t}\eta_{1i} + \gamma_{2t} x_{ti} + \varepsilon_{ti}$	*b*	- Person-centered method - Identification of different trajectories within population - Assigning individuals to a trajectory - Possibility to study each trajectory shape with static covariates - Possible consideration of cluster effect	- Questioning of the validity of individual trajectories caused by data missingness - Need for sufficient amount of data for each individual
GMM	50	$y_i = P_y \eta_i + \varepsilon_i$			
LCA	154	$P(Y = y) = \sum_{i=1}^{C} P(X = x) P(Y = y \mid X = x)$			
GBTM	78	$P(Y_j) = \sum_j \pi_j(x_i) P^j(Y_j)$			
COMPLEMENTARY APPROACHES					
SEM	380	$\eta = \beta\eta + \gamma\xi$	*a*	- Description of latent process	- Need for a large sample size to fit

Table A1. *Cont.*

Model	Search Strategy	Mathematical Formulation	Missing Data	Advantages	Drawbacks
CLPM	105	$x_{it} = \mu_t + p_{it}$ $y_{it} = \eta_t + q_{it}$	*b*	- Study of causal relationships	- Need for a large sample size to fit - Strong assumptions (stationarity and synchronicity)

Legend: Missing data: *a*: complete-cases analysis, *b*: possible estimates computation despite missing data, *: unbiased estimates. Number of articles (search strategy in PubMed): The keywords ("occupation*" OR "profession*" OR "job-related" OR "work-related") were linked with the following keywords: MLM: (multilevel OR multi-level). Repeated measures ANOVA: (ANOVA). Mixed model: (mixed). GCM: (GCM OR growth curve). GMM: (GMM OR growth mixture). LCA: (LCA OR LCGA OR latent class). GBTM: (GBTM or group-based trajector*). SEM: (SEM or structural equation). CLPM: (CLPM or cross-lagged).

Appendix B. Details for the Search Strategy Used within Each Database

Results of different research conducted on the Cochrane Library/Embase/PsycInfo/PubMed databases:

Cochrane Library

#1 occupational OR professional OR job-related OR work-related
#2 cluster
#3 missing data OR missing values
#4 longitudinal
#5 #1 AND #2 AND #3 AND #4
Filter Language = none in CENTRAL
Filter Dates = none
Results = 0

Embase

'occupational' OR 'professional' OR 'job-related' OR 'work-related' AND ('cluster*') AND ('missing data' OR 'missing value*') AND 'longitudinal '
Filter Language = none
Filter Dates = none
Results = 2

PsycInfo

Any Field: occupational *OR* **Any Field**: professional *OR* **Any Field**: job-related *OR* **Any Field**: work-related *AND* **Any Field**: cluster *AND* **Any Field**: missing data *OR* **Any Field**: missing values *AND* **Any Field**: longitudinal
Filter Language = none
Filter Dates = none
Results = 2

PubMed

("occupation*" OR "profession*" OR "job-related" OR "work-related") AND "cluster*" AND ("missing data" OR "missing value*") AND "longitudinal"
Filter Language = none
Filter Dates = none
Results = 3

References

1. Basch, E.; Schrag, D. The Evolving Uses of "Real-World" Data. *JAMA* **2019**, *321*, 1359–1360. [CrossRef] [PubMed]
2. Makady, A.; de Boer, A.; Hillege, H.; Klungel, O.; Goettsch, W. What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. *Value Health* **2017**, *20*, 858–865. [CrossRef] [PubMed]
3. Corrigan-Curay, J.; Sacks, L.; Woodcock, J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA* **2018**, *320*, 867–868. [CrossRef]
4. McCormick, J.; Mehta, G.; Olesen, H.V.; Viviani, L.; Macek, M.; Mehta, A. Comparative demographics of the European cystic fibrosis population: A cross-sectional database analysis. *Lancet* **2010**, *375*, 1007–1013. [CrossRef]

5. Dutheil, F.; Duclos, M.; Naughton, G.; Dewavrin, S.; Cornet, T.; Huguet, P.; Chatard, J.-C.; Pereira, B. Wittyfit-live your work differently: Study protocol for a workplace-delivered health promotion. *JMIR Res. Protoc.* **2017**, *6*, e6267. [CrossRef] [PubMed]
6. Platt, R.; Brown, J.S.; Robb, M.; McClellan, M.; Ball, R.; Nguyen, M.D.; Sherman, R.E. The FDA Sentinel Initiative—An Evolving National Resource. *N. Engl. J. Med.* **2018**, *379*, 2091–2093. [CrossRef]
7. Smith, C.A.; Wicks, P.J. PatientsLikeMe: Consumer Health Vocabulary as a Folksonomy. *AMIA Annu. Symp. Proc.* **2008**, *2008*, 682–686.
8. Randhawa, G.S. Building electronic data infrastructure for comparative effectiveness research: Accomplishments, lessons learned and future steps. *J. Comp. Eff. Res.* **2014**, *3*, 567–572. [CrossRef]
9. James, G.; Nyman, E.; Fitz-Randolph, M.; Niklasson, A.; Hedman, K.; Hedberg, J.; Wittbrodt, E.T.; Medin, J.; Moreno Quinn, C.; Allum, A.M.; et al. Characteristics, symptom severity, and experiences of patients reporting chronic kidney disease in the patientslikeme online health community: Retrospective and qualitative study. *J. Med. Internet Res.* **2020**, *22*, e18548. [CrossRef]
10. Benjdir, M.; Audureau, É.; Beresniak, A.; Coll, P.; Epaud, R.; Fiedler, K.; Jacquemin, B.; Niddam, L.; Pandis, S.N.; Pohlmann, G.; et al. Assessing the impact of exposome on the course of chronic obstructive pulmonary disease and cystc fibrosis: The REMEDIA European Project Approach. *Environ. Epidemiol.* **2021**, *5*, e165. [CrossRef]
11. McCaffrey, S.; Black, R.A.; Nagao, M.; Sepassi, M.; Sharma, G.; Thornton, S.; Kim, Y.H.; Braverman, J. Measurement of quality of life in patients with mycosis fungoides/sézary syndrome cutaneous t-cell lymphoma: Development of an electronic instrument. *J. Med. Internet Res.* **2019**, *21*, e11302. [CrossRef] [PubMed]
12. Maissenhaelter, B.E.; Woolmore, A.L.; Schlag, P.M. Real-world evidence research based on big data. *Onkologe* **2018**, *24* (Suppl. S2), 91–98. [CrossRef]
13. Garrison, L.P.; Neumann, P.J.; Erickson, P.; Marshall, D.; Mullins, C.D. Using Real-World Data for Coverage and Payment Decisions: The ISPOR Real-World Data Task Force Report. *Value Health* **2007**, *10*, 326–335. [CrossRef] [PubMed]
14. Barrett, J.S.; Heaton, P.M. Real-World Data: An Unrealized Opportunity in Global Health? *Clin. Pharmacol. Ther.* **2019**, *106*, 57–59. [CrossRef] [PubMed]
15. Han, J.; Pei, J.; Kamber, M. *Data Mining: Concepts and Techniques*, 3rd ed.; Elsevier: Amsterdam, The Netherlands, 2011; 740p.
16. Diggle, P.; Heagerty, P.; Liang, K.-Y.; Zeger, S. *Analysis of Longitudinal Data*, 2nd ed.; OUP: Oxford, UK, 2002; 396p.
17. Fitzmaurice, G.M.; Laird, N.M.; Ware, J.H. *Applied Longitudinal Analysis*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2012; 742p.
18. Caruana, E.J.; Roman, M.; Hernández-Sánchez, J.; Solli, P. Longitudinal studies. *J. Thorac. Dis.* **2015**, *7*, E537–E540. [PubMed]
19. Van Belle, G.; Fisher, L.D.; Heagerty, P.J.; Lumley, T. *Biostatistics: A Methodology for the Health Sciences*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2004; 895p.
20. Edwards, L.J. Modern statistical techniques for the analysis of longitudinal data in biomedical research. *Pediatr. Pulmonol.* **2000**, *30*, 330–344. [CrossRef]
21. Weiss, R.E. *Modeling Longitudinal Data*; Springer Science & Business Media: New York, NY, USA, 2005; 445p.
22. Killip, S.; Mahfoud, Z.; Pearce, K. What Is an Intracluster Correlation Coefficient? Crucial Concepts for Primary Care Researchers. *Ann. Fam. Med.* **2004**, *2*, 204–208. [CrossRef] [PubMed]
23. Song, P.X.-K. *Correlated Data Analysis: Modeling, Analytics, and Applications*; Springer Science & Business Media: New York, NY, USA, 2007; 356p.
24. Goldstein, H. *Multilevel Statistical Models*, 4th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2011; 376p.
25. Bliese, P.D.; Hanges, P.J. Being Both Too Liberal and Too Conservative: The Perils of Treating Grouped Data as though They Were Independent. *Organ. Res. Methods* **2004**, *7*, 400–417. [CrossRef]
26. Hayes, A.F. A Primer on Multilevel Modeling. *Hum. Commun. Res.* **2006**, *32*, 385–410. [CrossRef]
27. Gibbons, R.D.; Hedeker, D.; DuToit, S. Advances in analysis of longitudinal data. *Annu. Rev. Clin. Psychol.* **2010**, *6*, 79–107. [CrossRef]
28. Murray, D.M. *Design and Analysis of Group-Randomized Trials*; Oxford University Press: Oxford, UK, 1998; 481p.
29. Snijders, T.A.B.; Bosker, R.J. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd ed.; SAGE: Thousand Oaks, CA, USA, 2011; 370p.
30. Begg, M.D.; Parides, M.K. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Stat. Med.* **2003**, *22*, 2591–2602. [CrossRef] [PubMed]
31. Bruckers, L.; Molenberghs, G.; Pulinx, B.; Hellenthal, F.; Schurink, G. Cluster analysis for repeated data with dropout: Sensitivity analysis using a distal event. *J. Biopharm. Stat.* **2018**, *28*, 983–1004. [CrossRef] [PubMed]
32. Hox, J.J.; Moerbeek, M.; van de Schoot, R. *Multilevel Analysis: Techniques and Applications*, 3rd ed.; Routledge: Abingdon, UK, 2017; 365p.
33. Raudenbush, S.W.; Bryk, A.S. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed.; SAGE: Thousand Oaks, CA, USA, 2002; 520p.
34. Graham, J.W. Missing Data Analysis: Making It Work in the Real World. *Annu. Rev. Psychol.* **2008**, *60*, 549–576. [CrossRef] [PubMed]
35. Little, T.D.; Lang, K.M.; Wu, W.; Rhemtulla, M. Missing Data. In *Developmental Psychopathology*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2016; pp. 1–37.
36. Hedeker, D.; Gibbons, R.D. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychol. Methods* **1997**, *2*, 64–78. [CrossRef]
37. Kang, H. The prevention and handling of the missing data. *Korean J. Anesthesiol.* **2013**, *64*, 402–406. [CrossRef]

38. Donner, A. The Relative Effectiveness of Procedures Commonly Used in Multiple Regression Analysis for Dealing with Missing Values. *Am. Stat.* **1982**, *36*, 378–381.
39. Newgard, C.D.; Lewis, R.J. Missing Data: How to Best Account for What Is Not Known. *JAMA* **2015**, *314*, 940–941. [CrossRef]
40. Li, P.; Stuart, E.A.; Allison, D.B. Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA* **2015**, *314*, 1966–1967. [CrossRef]
41. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2002.
42. Allison, P.D. *Missing Data*; Quantitative Applications in the Social Sciences; SAGE Publications: Thousand Oaks, CA, USA, 2001; Volume 136.
43. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
44. Kenward, M.G.; Carpenter, J. Multiple imputation: Current perspectives. *Stat. Methods Med. Res.* **2007**, *16*, 199–218. [CrossRef]
45. Diggle, P.; Kenward, M.G. Informative Drop-Out in Longitudinal Data Analysis. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1994**, *43*, 49–73. [CrossRef]
46. Little, R.J.A. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *J. Am. Stat. Assoc.* **1995**, *90*, 1112–1121. [CrossRef]
47. Twisk, J.; de Vente, W. Attrition in longitudinal studies: How to deal with missing data. *J. Clin. Epidemiol.* **2002**, *55*, 329–337. [CrossRef]
48. Fitzmaurice, G. Missing data: Implications for analysis. *Nutrition* **2008**, *24*, 200–202. [CrossRef] [PubMed]
49. Rosenthal, S. Data Imputation. In *The International Encyclopedia of Communication Research Methods*; American Cancer Society: Atlanta, GA, USA, 2017; pp. 1–12.
50. Liu, C.; Cripe, T.P.; Kim, M.-O. Statistical Issues in Longitudinal Data Analysis for Treatment Efficacy Studies in the Biomedical Sciences. *Mol. Ther.* **2010**, *18*, 1724–1730. [CrossRef] [PubMed]
51. Verbeke, G. Linear Mixed Models for Longitudinal Data. In *Linear Mixed Models in Practice: A SAS-Oriented Approach*; Lecture Notes in Statistics; Verbeke, G., Molenberghs, G., Eds.; Springer: New York, NY, USA, 1997; pp. 63–153.
52. Verbeke, G.; Molenberghs, G. *Linear Mixed Models for Longitudinal Data*; Springer: New York, NY, USA, 2000.
53. Laird, N.M.; Ware, J.H. Random-effects models for longitudinal data. *Biometrics* **1982**, *38*, 963–974. [CrossRef] [PubMed]
54. Fahrmeir, L.; Tutz, G. *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed.; Springer Science & Business Media: New York, NY, USA, 1994; 537p.
55. Cnaan, A.; Laird, N.M.; Slasor, P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Stat. Med.* **1997**, *16*, 2349–2380. [CrossRef]
56. McCulloch, C.E.; Neuhaus, J.M. Generalized Linear Mixed Models. In *Encyclopedia of Biostatistics*; American Cancer Society: Atlanta, GA, USA, 2005.
57. Ju, K.; Lin, L.; Chu, H.; Cheng, L.-L.; Xu, C. Laplace approximation, penalized quasi-likelihood, and adaptive Gauss–Hermite quadrature for generalized linear mixed models: Towards meta-analysis of binary outcome with sparse data. *BMC Med. Res. Methodol.* **2020**, *20*, 152. [CrossRef]
58. Liang, K.-Y.; Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–22. [CrossRef]
59. Ballinger, G.A. Using generalized estimating equations for longitudinal data analysis. *Organ. Res. Methods* **2004**, *7*, 127–150. [CrossRef]
60. Zorn, C.J.W. Generalized estimating equation models for correlated data: A review with applications. *Am. J. Political Sci.* **2001**, *45*, 470–490. [CrossRef]
61. Bentler, P.M.; Weeks, D.G. Linear structural equations with latent variables. *Psychometrika* **1980**, *45*, 289–308. [CrossRef]
62. Hoyle, R.H. *Structural Equation Modeling: Concepts, Issues, and Applications*; SAGE: Thousand Oaks, CA, USA, 1995; 313p.
63. Ullman, J.B. Structural equation modeling: Reviewing the basics and moving forward. *J. Pers. Assess* **2006**, *87*, 35–50. [CrossRef] [PubMed]
64. Savalei, V.; Bentler, P.M. Structural Equation Modeling. In *The Corsini Encyclopedia of Psychology*; American Cancer Society: Atlanta, GA, USA, 2010; pp. 1–3.
65. Ullman, J.B.; Bentler, P.M. Structural Equation Modeling. In *Handbook of Psychology*, 2nd ed; American Cancer Society: Atlanta, GA, USA, 2012.
66. Kenny, D.A. Cross-lagged panel correlation: A test for spuriousness. *Psychol. Bull.* **1975**, *82*, 887–903. [CrossRef]
67. Selig, J.P.; Little, T.D. Autoregressive and cross-lagged panel analysis for longitudinal data. In *Handbook of Developmental Research Methods*; The Guilford Press: New York, NY, USA, 2012; pp. 265–278.
68. Kenny, D.A.; Harackiewicz, J.M. Cross-lagged panel correlation: Practice and promise. *J. Appl. Psychol.* **1979**, *64*, 372–379. [CrossRef]
69. Hamaker, E.L.; Kuiper, R.M.; Grasman, R.P.P.P. A critique of the cross-lagged panel model. *Psychol. Methods* **2015**, *20*, 102–116. [CrossRef]
70. Curran, P.J.; Willoughby, M.T. Implications of latent trajectory models for the study of developmental psychopathology. *Dev. Psychopathol.* **2003**, *15*, 581–612. [CrossRef]
71. Schumacker, R.; Lomax, R. *A Beginner's Guide to Structural Equation Modeling*, 4th ed.; Routledge: Mahwah, NJ, USA, 2016; Volume 288.
72. Muthén, B.; Muthén, L.K. Integrating Person-Centered and Variable-Centered Analyses: Growth Mixture Modeling with Latent Trajectory Classes. *Alcohol. Clin. Exp. Res.* **2000**, *24*, 882–891. [CrossRef]

73. Herle, M.; Micali, N.; Abdulkadir, M.; Loos, R.; Bryant-Waugh, R.; Hübel, C.; Bulik, C.M.; De Stavola, B.L. Identifying typical trajectories in longitudinal data: Modelling strategies and interpretations. *Eur. J. Epidemiol.* **2020**, *35*, 205–222. [CrossRef]
74. Nguena Nguefack, H.L.; Pagé, M.G.; Katz, J.; Choinière, M.; Vanasse, A.; Dorais, M.; Samb, O.M.; Lacasse, A. Trajectory Modelling Techniques Useful to Epidemiological Research: A Comparative Narrative Review of Approaches. *Clin. Epidemiol.* **2020**, *12*, 1205–1222. [CrossRef]
75. Hox, J.; Stoel, R.D. Multilevel and SEM approaches to growth curve modeling. In *Encyclopedia of Statistics in Behavioral Science*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2005.
76. Muthén, B.; Shedden, K. Finite Mixture Modeling with Mixture Outcomes Using the EM Algorithm. *Biometrics* **1999**, *55*, 463–469. [CrossRef] [PubMed]
77. Muthén, B. Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class–latent growth modeling. In *New Methods for the Analysis of Change*; American Psychological Association: Washington, DC, USA, 2001; pp. 291–322.
78. Nagin, D.S. Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychol. Methods* **1999**, *4*, 139–157. [CrossRef]
79. Nagin, D.S. *Group-Based Modeling of Development*; Harvard University Press: Cambridge, MA, USA, 2005; 226p.
80. Nagin, D.S.; Odgers, C.L. Group-based trajectory modeling in clinical research. *Annu. Rev. Clin. Psychol.* **2010**, *6*, 109–138. [CrossRef] [PubMed]
81. Nagin, D.S.; Jones, B.L.; Passos, V.L.; Tremblay, R.E. Group-based multi-trajectory modeling. *Stat. Methods Med. Res.* **2018**, *27*, 2015–2023. [CrossRef] [PubMed]
82. Lanza, S.T.; Rhoades, B.L. Latent Class Analysis: An Alternative Perspective on Subgroup Analysis in Prevention and Treatment. *Prev. Sci.* **2013**, *14*, 157–168. [CrossRef]
83. Lanza, S.T.; Cooper, B.R. Latent Class Analysis for Developmental Research. *Child Dev. Perspect.* **2016**, *10*, 59–64. [CrossRef]
84. Dupéré, V.; Lacourse, E.; Vitaro, F.; Tremblay, R.E. Méthodes d'analyse du changement fondées sur les trajectoires de développement individuel. Modèles de régression mixtes paramétriques et non paramétriques. *Bull. Méthodol. Sociol. Bull. Sociol. Methodol.* **2007**, *95*, 26–57. [CrossRef]
85. Rogosa, D.; Brandt, D.; Zimowski, M. A growth curve approach to the measurement of change. *Psychol. Bull.* **1982**, *92*, 726–748. [CrossRef]
86. Martin, D.P.; von Oertzen, T. Growth mixture models outperform simpler clustering algorithms when detecting longitudinal heterogeneity, even with small sample sizes. *Struct. Equ. Model. A Multidiscip. J.* **2015**, *22*, 264–275. [CrossRef]
87. McNeish, D.; Harring, J.R. The effect of model misspecification on growth mixture model class enumeration. *J. Classif.* **2017**, *34*, 223–248. [CrossRef]
88. McNeish, D.; Matta, T. Differentiating between mixed-effects and latent-curve approaches to growth modeling. *Behav. Res.* **2018**, *50*, 1398–1414. [CrossRef] [PubMed]
89. Den Teuling, N.G.P.; Pauws, S.C.; van den Heuvel, E.R. A comparison of methods for clustering longitudinal data with slowly changing trends. *Commun. Stat.-Simul. Comput.* **2021**, *20*, 1–28. [CrossRef]
90. Nelder, J.A.; Wedderburn, R.W.M. Generalized Linear Models. *J. R. Stat. Soc. Ser. A (Gen.)* **1972**, *135*, 370–384. [CrossRef]
91. Booth, J.G.; Hobert, J.P. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **1999**, *61*, 265–285. [CrossRef]
92. Shapiro, A.; Browne, M.W. Analysis of covariance structures under elliptical distributions. *J. Am. Stat. Assoc.* **1987**, *82*, 1092–1097. [CrossRef]
93. Browne, M.W. Asymptotically distribution-free methods for the analysis of covariance structures. *Br. J. Math. Stat. Psychol.* **1984**, *37*, 62–83. [CrossRef] [PubMed]
94. Allison, P.D.; Williams, R.; Moral-Benito, E. Maximum likelihood for cross-lagged panel models with fixed effects. *Socius* **2017**, *3*, 1–17. [CrossRef]
95. Zyphur, M.J.; Hamaker, E.L.; Tay, L.; Voelkle, M.; Preacher, K.J.; Zhang, Z.; Allison, P.D.; Pierides, D.C.; Koval, P.; Diener, E.F. From data to causes III: Bayesian priors for general cross-lagged panel models (GCLM). *Front. Psychol.* **2021**, *12*, 612251. [CrossRef]